


Tailored IoT & BigData Sandboxes and Testbeds for Smart,  
Autonomous and Personalized Services in the European  
Finance and Insurance Services Ecosystem

# Infinitech

## D3.12 – Data Governance Framework and Tools - I

<b>Revision Number</b>	3.0
<b>Task Reference</b>	T3.5
<b>Lead Beneficiary</b>	GRAD
<b>Responsible</b>	Lilian Adkinson Orellana
<b>Partners</b>	ATOS, JSI, DWF, GRAD
<b>Deliverable Type</b>	Report (R)
<b>Dissemination Level</b>	Public (PU)
<b>Due Date</b>	2020-11-30
<b>Delivered Date</b>	2020-11-30
<b>Internal Reviewers</b>	BOUN, CCA
<b>Quality Assurance</b>	INNOV
<b>Acceptance</b>	WP Leader Accepted and/or Coordinator Accepted
<b>EC Project Officer</b>	Pierre-Paul Sondag
<b>Programme</b>	HORIZON 2020 - ICT-11-2018
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no 856632

## Contributing Partners

Partner Acronym	Role	Author(s)
GRAD	Lead Beneficiary	Lilian Adkinson Orellana Marta Sestelo Pérez Borja Pintos Castro
ATOS	Contributor	Nuria Ituarte Aranda Darío Ruíz López
JSI	Contributor	Maja Skrjanc
BOUN	Internal Reviewer	Can Özturan
CCA	Internal Reviewer	Paul Lefrere
INNOV	Quality Assurance	John Soldatos

## Revision History

Version	Date	Partner(s)	Description
0.1	2020-04-06	GRAD	ToC Version
1.0	2020-11-13	GRAD, ATOS, JSI	First Version for Internal Review
2.0	2020-11-24	GRAD	Version for Quality Assurance
3.0	2020-11-30	GRAD	Version for Submission

## Executive Summary

This document determines how Data Governance Framework and Tools will be carried out and developed through the INFINITECH project in order to ensure the appropriate behaviour of data and analytics. Data Governance is an essential axis that guarantees data security and establishes which is the route to follow in the management of information within all companies, and particularly in banks, Fintechs and other financial organizations. In order to do this, the INFINITECH project will provide the following tools and mechanisms:

- i) **A pseudonymization tool;**
- ii) **A tool for anonymizing data;**
- iii) **A set of mobile digital user onboarding services, and**
- iv) **A solution for authenticating citizens and/or businesses against the eIDAS infrastructure.**

The first tool will afford pseudonymization of unique identifiers and generalization of numeric and time-stamped fields in batch mode by means of different techniques. The second tool for anonymized data will determine automatically the best configuration for each application in such a way that individuals' data remain anonymous. The third solution corresponds to a mobile digital user onboarding service that will be developed. It will allow for remote user registration using eID or electronic password and will provide multi-factor authentication combining images of the face of the user with the certificates stored in the eID or passport. Finally, the last tool will provide a cross-border strong authentication mechanism based on eIDs and will support authentication for citizens, compatibility with the eIDAS Network, and use of eIDs issued by European National authorities according to the EU eID schemas.

The deliverable starts by reviewing the state of the art about all the techniques used in each of the tools described above. Then, the preliminary design of these tools is defined and explained, and finally, a set of conclusions is included in order to summarize the most important reviewed topics.

## Table of Contents

<b>1 Introduction</b>	6
1.1 Objective of the Deliverable	6
1.2 Insights from other Tasks and Deliverables	6
1.3 Structure	7
<b>2 State of the art of data governance technologies</b>	8
2.1 Review on pseudonymization mechanisms	8
2.2 Review on data anonymization mechanisms	9
2.3 Review on digital user onboarding services	11
2.4 Review on authentication in eIDAS infrastructure	13
<b>3 Preliminary design of the data governance framework and tools</b>	15
3.1 Data pseudonymization tool	15
3.2 Data anonymization	16
3.3 Digital user onboarding services tool	19
3.4 Authentication tool for the eIDAS infrastructure	20
<b>4 Conclusions</b>	22
<b>Appendix A: Literature</b>	23

## List of Figures

Figure 1 - eIDAS network schema	14
Figure 2 - Schema of the pseudonymization tool	16
Figure 3 - Obtaining of the configuration for anonymizing a dataset	17
Figure 4 - Operation of the anonymization service in <i>batch mode</i>	18
Figure 5 - Operation of the anonymization service in <i>streaming mode</i>	18
Figure 6 - Configuration of the streaming mode of the anonymization service	19
Figure 7 - Overall work procedure of DUOS	19
Figure 8 - INFINITECH Onboarding operation	20
Figure 9 - INFINITECH authentication against vID generated in the onboarding process	20
Figure 10 - Spanish citizen accessing online service provided by a company based in the EU and connected to eIDAS network	21

## Abbreviations/Acronyms

Abbreviation    Definition

4AML	Fourth Money Laundering Directive
AI/ML	Artificial Intelligence/Machine Learning
API	Application Programming Interfaces
ARIES	reliAble euRopean Identity EcoSystem
CEF	Connecting Europe Facility
DPO	Data Protection Orchestrator
DUOS	Digital User Onboarding System
eDNI	National Electronic Identifier
eID	Electronic Identification
eIDAS	Electronic Identity And trust Services
GDPR	General Data Protection Regulation
GPS	Global Positioning System
HPC	High Performance Computing
IdP	Identity provider
IMDb	Internet Movie Database
IoT	Internet of Things
JWT	JSON Web Token
LEPS	Leveraging eID in the Private Sector
MAC	Media Access Control
MD5	Message-Digest Algorithm 5
MiFID	Markets in Financial Instruments Directive
MS	Member States
PIAM	Platforms for identity and access management
PKI	Public key infrastructure
PSD	Payment Service Provider
RA	Reference Architecture
REST	Representational State Transfer
RNG	Random Number Generator
SAML	Security Assertion Markup Language
SHA	Secure Hash Algorithm
SHAPR	Smart, Holistic, Autonomy, Personalized and Regulatory Compliance
SME	Small and medium-sized enterprises
SP	Supports Public
SPeIDI	Service Provider for eIDas Integration
SW	Software
URL	Uniform Resource Identifier
VID	Virtual Identification

# 1 Introduction

This document is the first deliverable of a series of three whose goal is to present the data governance mechanisms that will be developed within the INFINITECH project during the 26 months of the task “T3.5 Data Governance Mechanisms”. In this first deliverable, a preliminary overview of these mechanisms is presented.

## 1.1 Objective of the Deliverable

The objective of this deliverable is twofold: on the one hand, it includes a review of the state of the art of the most common data governance mechanisms, including the technologies listed below as well as other relevant techniques; on the other hand, it presents a preliminary design of four tools that will be developed in the scope of the task. Specifically, these mechanisms are the following:

1. A **pseudonymization tool** that determines automatically the best anonymization configuration for each application.
2. A **tool for anonymizing data** that determines automatically the best anonymization configuration for each application.
3. A **mobile digital user onboarding services** with virtual eID derived from government issued documents (ePassport or eID card).
4. A **solution for authenticating** citizens and businesses **against the eIDAS infrastructure**, providing a cross-border strong authentication mechanism based on eIDs.

The four tools that will be developed within this task are a relevant piece in banks, Fintechs and other financial organizations, and this is why Task 3.5 is included in the INFINITECH project. It is worth mentioning that data in financial services need to be anonymized or pseudonymized, users of digital financial services (e.g., retail investors, SME owners, etc.) need to be digitally on boarded to the services and that the connection of online services with eIDAS infrastructure allows the authentication of customers against this pan European eID infrastructure ensuring the legal validity of cross-border transactions.

## 1.2 Insights from other Tasks and Deliverables

The work that is presented in this deliverable is based on the corresponding task “T3.5 Data Governance Mechanisms”, which is included in the “WP3 BigData/IoT Data Management and Governance for SHARP Services”. The deliverable is widely related to other tasks and deliverables. Specifically:

- **D2.5 Specifications of INFINITECH Technologies - I.** That deliverable lists the tools and technologies currently available and in development by the partners of INFINITECH. Additionally, it contains the specifications of the components identified by the pilots, including Input and Output formats, functionalities and specifications about the implementation technologies (e.g. BigData/IoT platforms, AI/ML toolkits, HPC infrastructures) that will be used to realize them. Within the proposed component groups, there is one related to Security and Privacy which includes eIDAS integration, Data Anonymization, Digital User Onboarding System (DUOS) and Data Protection Orchestrator (DPO).
- **D2.7 Security and Regulatory Compliance Specifications - I.** The aim of that deliverable is the specification of the standards and regulatory environment of the INFINITECH project. An important emphasis is put on the GDPR due to its high relevance in BigData and data analytics frameworks which apply to INFINITECH’s sharp services. Moreover, regulations such as PSD II, MiFiD II and 4AMLD which are relevant in the Financial Sector are assessed with respect to the INFINITECH pilot scenarios.
- **D2.13 Reference Architecture - I.** That deliverable presents the first version of the INFINITECH Reference Architecture (RA). This RA provides a schema for building solid workflows and ensures full

communication and interaction between all the building blocks which include the integration of the Data Governance Tools.

- **INFINITECH D2.15 Regulatory Compliance Tools - I.** This deliverable analyses regulatory compliance through the INFINITECH project, specifically in every pilot. The main relation with the present deliverable is that the regulatory compliance tools implemented in T3.6 will call to the components developed here.

The data governance mechanisms that are being developed within this task will be applied and validated in the pilots of “WP7 Large-Scale Pilots of SHARP Financial and Insurance Services”, according to the issues and regulations to fulfil.

## 1.3 Structure

The structure of this document is as follows. Section 1 contains the introduction of the document, including its objectives and the relation with other tasks and deliverables of the project. In section 2, a review of the state of the art of data governance mechanisms is presented, including technologies such as pseudonymization, data anonymization, authentication mechanisms and digital user onboarding techniques, as well as a compilation of other data governance technologies.

Section 3 contains a description of the preliminary design of the tools that are going to be developed in the task, i.e, a pseudonymization tool, a data anonymization tool, a digital user onboarding service and an authentication tool for the eIDAS infrastructure. Finally, section 4 concludes the document summarizing the most relevant concepts.

## 2 State of the art of data governance technologies

The present section tries to review and summarize the state of the art of the data governance technologies that will be developed in the INFINITECH project. Firstly, an overview of the pseudonymization mechanisms is presented, including counter and random number generators, cryptographic hash functions, symmetrical encryption techniques or data masking. Secondly, data anonymization methodologies are exposed, both for general purpose data as well as others that focus on geo-located data. eIDAS (electronic Identity And trust Services) Regulation and the process of digital user onboarding are explained in detail in the last two subsections.

### 2.1 Review on pseudonymization mechanisms

When working with sensitive data, such as healthcare or financial transaction data, care should be taken regarding data protection, which is highly regulated. GDPR, as one of the key valid regulations dealing with data, permits processing of personal data for a purpose other than originally intended, in “the existence of appropriate safeguards, which may include encryption or pseudonymization.” Other purposes can include profiling, business analysis, outsourcing data processing to non EU/EEA countries, and use for scientific, historical, and statistical purposes.

GDPR also makes pseudonymization a central feature of the requirement for data protection by design and by default, and enables processing personal data for scientific, historical, and statistical purposes if the data is safeguarded by pseudonymization. Therefore, pseudonymization can be used when realistic data is needed for application development and testing environments, data warehousing, analytical data stores, training programs, or other business processes. It can also be used when exporting data to non-EU/EEA countries.

**Pseudonymization** is a security technique for **replacing sensitive data with realistic fictional data that cannot be attributed to a specific individual without additional information** which, according to GDPR, is to be “kept separately and subject to technical and organisation measures to ensure non-attribution to an identified or identifiable person” [1]. The technique allows to maintain referential integrity and statistical accuracy, thereby enabling business processes, development and testing systems, training programs, and analysis to operate normally.

In order to ensure privacy, several pseudonymization methods can be applied to the data to either pseudonymize direct identifiers (e.g. Social Security Numbers) or to scramble the data. As an example, person-specific ID’s should in most scenarios be replaced by pseudonyms (hence the name pseudo-anonymization or pseudonymization), making third parties unable to convert pseudonyms back to the identifiers. However, records of an individual are still linked together with the pseudonym [1].

There are numerous ways of generating the pseudonyms out of raw identifiers [1]. **Counter and random number generator (RNG)** are the simplest pseudonymization functions. Pseudonyms generated using these functions are numbers and should be stored together with the original identifiers at a central, private location. Alternatively, applying a **cryptographic hash function** to the identifier is generally considered a weak pseudonymization function as it is prone to brute force and dictionary attacks. A secret key that is stored together with the original data can be added to the input in order to generate more secure pseudonyms. Without a knowledge of this key it is not possible to map the identifiers to pseudonyms. Finally, **symmetrically encrypted identifiers** may also be used as pseudonyms. This approach is similar to the previous one. Deterministic or probabilistic encryption may be chosen based on the implementation. Alternatively, **data masking** replaces sensitive data with fictitious yet realistic data, which helps reduce data risk while preserving data utility.

Moreover, while choosing the pseudonymization function determines what the pseudonyms will look like, the **policy of implementation** deals with the degree of pseudonymization. It is possible, for example, to use the same pseudonym for each identifier (**deterministic pseudonymization**), to only use the same pseudonyms for an identifier appearing across different databases but not in the same database (**document-randomized pseudonymization**), or to use a different pseudonym for each entry (**fully randomized**



**pseudonymization**) [1]. However, the appropriate policy should be chosen depending on its practical application [2].

It is important to highlight that the GDPR both encourages pseudonymization and distinguishes it from anonymous data. When applying pseudonymization, the individual can be identified by linking pseudonymized and additional non-pseudonymized information (e.g., birth date, gender, zip code). To address the fact that pseudonymized data is not anonymous, the GDPR requires the following:

- pseudonymized data to be treated as personal data if a specific individual can be identified “by the use of additional information.” As such, appropriate and effective technological and organization measures must be implemented to protect the pseudonymized data.
- pseudonymized and “additional information for attributing the personal data to a specific data subject” to be kept separate.
- implementing appropriate technical safeguards (e.g., encryption, hashing, or tokenization) and organizational policies to prevent unauthorized reversal of pseudonymization.

Sometimes, re-identification can be possible even without direct identifiers, as the data contains also other attributes that, if linked to an external database, could allow recovering the original identities. For example, by knowing an individual’s zip code and their rare medical condition one might be able to recover the patient’s full identity (if, perhaps, the patient is the only person to have that medical condition within the area). In the financial sector there are different types of information which are considered sensitive data. For instance, in case of transaction data, besides payee and payer information, one must camouflage also the exact amount of transaction. Therefore, depending on the concrete application, pseudonymization or anonymization techniques can be applied to sensitive data that will be released for research purposes, in order to limit disclosure risk while trying to maximize the amount of information in the data [3].

## 2.2 Review of data anonymization mechanisms

The term “privacy” covers a wide range of concepts and definitions: bodily privacy, which means that your body is your own and it is related to protection from physically invasive procedures, such as genetic testing; territorial privacy, which concerns the setting of limits on intrusions into physical space, such as companies or homes; communication privacy, focused on the security of communications, such as email or messages through WhatsApp; and information privacy, which deals with the establishment of rules governing the collection, processing and handling of personal data [4]. With the focus on this last term, **data anonymization** (or de-identification) is intended to **process personal data in order to irreversibly prevent identification** [5]. The idea behind this approach to handling is that the information that the adversary can use in order to link the data with an individual is removed or transformed in such a way that the observable data cannot be used to breach users’ privacy. Citing the General Data Protection Regulation (GDPR) [5], “the principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person”, so once the data is truly anonymized those principles will no longer apply. The problem is that applying data anonymization correctly is a challenging task and it is necessary to assess its success, typically by measuring an individual’s risk of re-identification [6], [7].

Recent examples have been announced in which it is confirmed that deleting the identifiable information of people included within the disclosed dataset is not sufficient in order to preserve their privacy. The adversary might have access to other auxiliary data that, once combined with the anonymized data, could lead to the re-identification of users or to the publication of some sensitive information about them. Netflix is one of the most well-known cases [8]. Two researchers of the University of Texas were able to re-identify some portion of anonymized Netflix movie-ranking data from individual consumers on the streaming website by comparing them with non-anonymous IMDb (Internet Movie Database) users’ movie ratings. The data was released by Netflix 2006 after de-identification, which consisted of replacing individual names with random numbers and moving around personal details. This kind of risk exists not only if the identifiers are removed but also if the data is pseudonymized, i.e., when substituting an identifier by a pseudonym (typically through encryption, hashing or tokenization). Also in 2006 AOL released 20 million search queries for 650,000 users, from three

months of data. The company attempted to suppress identifying information, including usernames and IP addresses, but, to preserve the data's utility, AOL replaced that information with unique identifying numbers through pseudonymization. Within days of the database's release, journalists from the New York Times had revealed the identity of user number 4417749 to be Thelma Arnold, a 62-year-old widow from Lilburn, Ga. The fallout from the AOL incident was devastating, both for the company and the industry as a whole. Therefore, in order to adequately address the anonymization processes it is necessary not only to anonymize the data but also to be able to estimate the risk of re-identification.

Most of the anonymization techniques are based on two main methods: **randomization** and **generalization**. They respectively consist of adding a certain randomness to numeric data values and “replacing a value with a less specific but semantically consistent value” [9]. Note that, conceptually, **noise addition** is the simplest way of randomizing data [10]. The idea behind this technique is to add statistical noise to a dataset but keeping the same distribution. Another option often used when the original distribution of data is required for processing purposes are **permutation techniques**, which involve shuffling the relationships of the datasets, linking a certain sensitive attribute to other individuals [11]. **Differential privacy** is another well-known anonymization technique that relies on variable noise as a means of ensuring that the risk incurred by participating in a dataset is only marginally greater than the risk of not participating in it [12].

Regarding the generalization of the information contained in the dataset, **k-anonymity**, *l*-diversity, *t*-closeness are other alternatives to anonymize data. In the first case, privacy is achieved by grouping attributes from at least *k* individuals [13]. ***l*-diversity** extends this idea by ensuring that each aggregated attribute will contain at least *l* different values [14]. Finally, ***t*-closeness** improves the previous methods by preserving the original data distribution, guaranteeing that each value of the aggregated attributes will appear in the anonymized data as many times as in the original dataset [15].

Coming back to the first paragraph of this section where classification about types of privacy were cited, location privacy can be defined as a special type of information privacy which covers the rights of individuals to determine for themselves when, how, and to what extent their location information can be known and processed. In short, the ability of an individual to control access to their current and past location information is the central issue in **location privacy**.

Nowadays, there are a high number of applications and architectures that will ask users to reveal their location. For instance, through the use of a mobile, a user could request from the network, bus schedules, movie times, local information, etc. based on the user's current location. If an attacker would exploit this location data, the risk on privacy could be easily compromised. Examples of studies in which location data is used to demonstrate a privacy attack can be found in the literature [16]–[19]. As a particular case, Hoh et al. [20] reference is exposed. In this work, the authors used a database of week-long GPS traces with information of 239 drivers in Detroit. Focusing on a subset of 65 drivers, their home-finding algorithm found plausible home locations for about 85% of them, although the authors did not know the actual locations of the drivers' homes. In more recent studies [21], it is shown that four randomly chosen points are enough to uniquely characterize the movements of the 95% of the users of a dataset, and with the selection of just two randomly chosen points it would still be possible to characterize more than 50% of them. Therefore, mobility traces can be considered in general as unique and thus, it cannot be stated that a dataset that only contains location data will be anonymous per se.

It is also important to highlight that, in addition to the use of location data to infer person's home and work location, it is possible to obtain other sensitive information as the moving person's mode of transportation (i.e., bus, foot, car) [22], their route's prediction [23] or the end of a trip [24]. Based on the previous points, it is clearly seen that location can be used to infer a lot of information about a person, even if the person's name is not included in the dataset. In fact, Bettini et al. [25] indicate that a set of recorded locations for a person constitutes a quasi-identifier, i.e., data that can be used in combination with other information in order to identify the user.

There exist many strategies proposed and tested for improving location privacy. According to Duckham and Kulik [26] some of them are related to regulatory strategies or privacy policies while others are based on anonymity or obfuscation. Note that this deliverable intends to review the later.

According to anonymization mechanisms, the notion of k-anonymity is the most widely used definition of privacy for location in the literature (see [27]–[29] among others). The method achieves privacy protection by means of generalization and suppression algorithms in order to ensure that any one user is indistinguishable from other users (k-1). Paying attention to geolocated data, a subject is considered k-anonymous if its location is indistinguishable from those of k-1 other users.

**Spatial cloaking**, first proposed in [27], is based on one of the most popular methods in the anonymity literature group-anonymity. It tries to make sure that a person will be indistinguishable from a group by means of a cloaked area large enough to contain the group size necessary to meet the intended anonymity constraint. An improvement of the adaptive spatial cloaking is proposed in [27]. The **combination of spatial cloaking with the temporal version** tries to reveal spatial coordinates with more accuracy while the accuracy in time is reduced. The desired spatial resolution is provided as an additional parameter in the spatio-temporal algorithm, and this determines the monitored area by dividing the space until the chosen resolution is obtained. Then, the algorithm registers the movements of the vehicles in these areas and delays the request to the Location Based Service until a number k minimum of vehicles have visited the area chosen for the requester. Afterward, time interval  $[t_1, t_2]$  is obtained by setting  $t_2$  to the current time, and  $t_1$  to the time of request minus a random cloaking factor. Finally, the area and the time interval are returned.

As we mentioned, another line of research related to location privacy protection is the use of **obfuscation mechanisms**. Obfuscation means the practice of willfully degrading the quality of information in some way in order to protect the privacy of the individuals from whom this information has been collected. Thus, location or spatial obfuscation is complementary to anonymity. In fact, rather than anonymizing users' identities, the solutions based on obfuscation introduce perturbations into the real locations to decrease their accuracy. Duckham and Kulik [30] worked out an obfuscation method for protecting geolocated data by artificially inserting into measurements some fake points with the same probability as the real user position. In the works of Cheng et al. [31] and Ardagna et al. [32] a method that sends an area dynamically calculated around the user's positions instead of the current position is proposed.

Finally, the concept of **differential privacy**, originally proposed in the area of statistical databases, is also applied to this framework. The idea is to protect the data of each individual in the dataset by means of publishing the information in an aggregated manner. With the focus on this, some controlled noise is added to the query output so that changing data of one user will have an undetectable effect on the given response. Most of the studies that apply this technique to location data have taken into consideration scenarios where aggregated information about a group of users is published and therefore, the application is the same as in other databases [33]–[35]. However, if we want to protect the location privacy of a single user the technique has limitations because it mainly requires data from a set of individuals. In any case, considering this single user and taking into account that by definition any change in their location should have negligible effect on the published response, the communication of any useful information to the service provider will be impossible; that is, privacy would be obtained but the utility will have a very big loss.

To solve this issue, Dewri [36] proposed to apply jointly both differential privacy and k-anonymity. In order to do this, they fix an anonymity set of k locations and establish as a requisite that the probability of giving the same obfuscated location z from any of these k locations should be similar (up to  $\epsilon$  which is established in the definition of the differential privacy method). Another approach is the notion of geo-indistinguishability [37], an instance of a generalized version of differential privacy, that guarantees that obfuscated locations within a radius around the real location of the user are statistically indistinguishable from other locations.

## 2.3 Review of digital user onboarding services

**Digital user onboarding is the process of enrolling new users for accessing services provided by an organization.** This process would be performed online, and the users would become a new customer in a remote and secure way. The reason for enrolling the users is that the system can remember them when they try to access the system, in order to check their identity together with the functionality and data that they are entitled to access. The process of checking the identity of the user, that is, that the person trying to access

the system is whom he claims to be receiving the name of **authentication**, and the process of checking what functionalities he is entitled to access is named **authorization or access control**. As there is no point in checking the functionalities that a given user is entitled to access without having checked its identity, authorization always relies on authentication. Moreover, authorization often checks additional information from the user, such as their age or very commonly, a role assigned to them. For this reason, the process of onboarding often involves retrieving and storing additional information from the user, so it can be used later in the system. This information associated with the user is named a digital identity, and the software that stores that information and provides it is named an **Identity provider (IdP)**.

Authentication, identity provision and authorization are complex but very common functionalities in any software system and led to the creation of software tools that deal with them. The most advanced tools are capable of carrying out all of these three functionalities. These tools are named **platforms for identity and access management (PIAM)**. As digital onboarding is closely related with authentication, many PIAM offer digital onboarding features, too. From now on, when we refer to IdPs, it is important to remember that their functions will often be provided by a PIAM.

Once the user becomes a customer of an organization, which in case of using an IdP implies that they are properly added to the list of users recognized by it and their information is properly stored in the IdP, they can be identified and given access to the services that they contracted or also they could also leave the services.

The way to perform digital user onboarding is through using virtual remote **electronic identification (eID)**. Virtual eID allows the companies to offer an online and remote process to perform digital user onboarding to the potential users allowing them to access their services. Virtual eID should be linked with a physical ID delivered by the government and also it would be convenient to perform multi-factor authentication to ensure the correct authorization of a person to access. The IdP usually allows several enrolments using different credentials such as national ID card as the first means of authentication and this credential can be linked with the other credentials with low level of assurance. The IdP is in charge of linking the credentials [38].

When using Virtual eIDs it is common to utilize PKI token, but it is also needed for the user information that is stored in the attributes which are always retrieved from the IdP. Currently, the state of the art in attributes retrieving has two options: extract them from the ID card or retrieve from the IdP. The level of assurance of the attributes is also stored also in the attributes source. There is also the possibility of retrieving certified attributes involving eIDAS solutions.

Current authentication systems using Virtual eID show the following weaknesses as stated in [38]:

1. The IdP is the center of the architecture storing information and managing policies. This makes the system weak in case of attack given that IdPs are usually a target for hackers.
2. Continuing with the centralized IdP component, the user shares all the information with the IdP and the privacy preservation depends on the IdP.
3. New virtual IDs have no link with real user identity. Electronic documents and biometric verification in the creation of new virtual IIDs should be used to ensure and verify the authentication and prevent from suspicious behaviours.
4. The attributes used in the enrolment are provided by the user. They do not come from the documents and this provides a low level of assurance.

Some of these weaknesses can be mitigated with the provision of third party authentication modules that complement the ones provided by the IdP, which is a feature supported by many PIAMs. This way a hacker would have to deceive all the authentication modules instead of only the central one. Additionally, third party authentication modules may provide additional authentication features that are not supported by a specific PIAM vendor, such as biometric checks mentioned in 3. A practice that is gaining momentum is that a biometric module, often provided by a third party, retrieves some biometric data from the user when enrolling in the system, for instance, taking a small video and storing some information retrieved from it. With that, an authentication module will be able to take another small video of the person trying to access

the system and compare the information retrieved from the video of the user accessing the system with the one stored from the person that was previously enrolled.

But the information used for authentication and therefore retrieved in the enrolment may not be limited to information of the user. It may also contain information that is related to the user but it is not about them. For instance, the enrolment process may also retrieve and store the MAC address of the system that the user is trying to use for getting enrolled, so the system can check later that it is being accessed from the same device. Therefore, a good example of multi-factor authentication could be to combine a national ID with a live video of the person taken from their own personal smartphone.

Finally, as some features of the authentication mechanisms are at risk of becoming obsolete, the enrolment process should provide the ability for users to update their information. For instance, if the system is using the MAC address of a user to identify them, then it should also allow them to change it.

## 2.4 Review of authentication in eIDAS infrastructure

eIDAS (electronic Identity And trust Services) Regulation is defined in EU 910/2014 [39]. The Regulation establishes:

- a legal framework for EU Digital Single Market in secure and cross-border transactions,
- a trust model for mutual recognition using e-identification means,
- eID and electronic Trust Services such as electronic signatures, electronic seals, time stamping, electronic registered delivery service and website authentication.

The goals of eIDAS regulation are to remove barriers, get an efficient digital single market increasing digital trust and increase legal assurance using the same rules across all Member States (MS). The regulation relies on a set of horizontal principles such as liability, supervision, international aspects, security requirements, Levels of Assurance, qualified services, trusted lists, technology neutrality, and EU trust mark.

eIDAS regulation contributes to the digital single market **allowing the mutual recognition of notified eID across Europe borders**. The eIDs schemes promoted by the Member States are secure authentication means that enable cross-border access to for example eGovernment services. Thereby natural & legal persons will authenticate to online services in any MS using their own national eIDs with security, ease of use and cost efficiency [40].

Member States may require the use of national eID(s) scheme(s) for, at least, access to public services through the eIDAS network. They must recognize and accept notified eIDs of other Member States for cross-border access to all the public services requiring e-identification [40]. Moreover, Member States must provide online free eID authentication and liability for identification of persons. They also must allow the private sector to use notified eID. Member States allow interoperability through eIDAS network using SAML 2.0 [41] and encryption to communicate with eIDAS nodes.

Regarding identity for eID schemes, eIDAS provides **three Levels of Assurance (LoA)**: Low, Substantial and High. They are based on ISO 29115 [42] and STORK QAA considering the practical experience gained during the STORK pilot [43].

The Connecting Europe Facility (CEF) programme supports Public SPs to offer online services capable of identifying citizens and businesses from other Member States and Private SPs to offer online services capable of identifying customers and consumers from other Member States. It provides help in defining technical specifications and sample SW [44].

eIDAS Network consists of several eIDAS nodes distributed in the countries. This scheme can be seen in Figure 1.

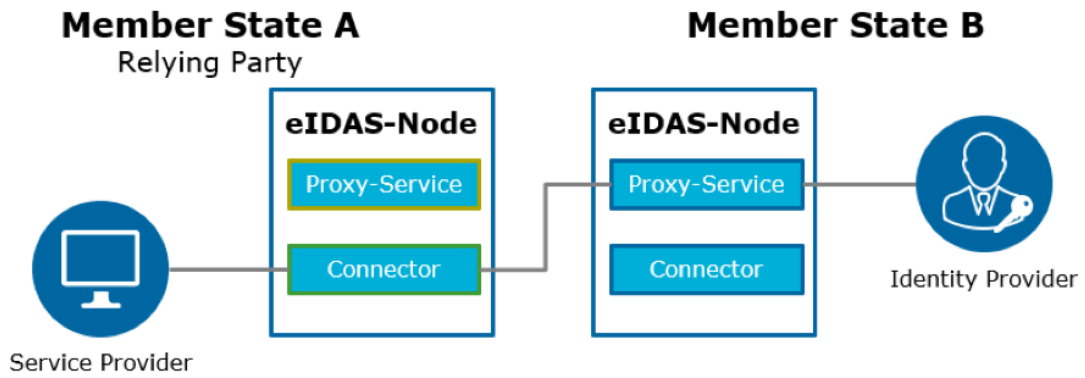


Figure 1 - eIDAS network schema.

In summary, as described in the LEPS project: “the eIDAS Network transports (across borders) the identity attributes that a Service Provider (SP) has requested for authenticating a user (for example, name, surname, country of origin, unique ID, etc.), from the authoritative source that hosts the eID of this user (also called Identity Provider or IdP) to the Service Providers IT infrastructure -- under the condition that the user has provided consent to such an information transfer” [45].

In the eIDAS network, the authentication mechanisms needed to authenticate a user and allow him/her to access the services from a service provider are given by the eIDAS network. This allows access to the services that are provided by the SPs using the eIDAS Network. That is, the eIDAS Network itself defines and deploys a pan-European infrastructure for cross-border user authentication allowing a user to utilize their national eID schema and use it to interconnect between other SPs and IdPs located in other European countries. The SP would be a Relying Party of the eIDAS network, and it will be connected to the eIDAS node using SAML2.0 [45].

## 3 Preliminary design of the data governance framework and tools

The present section is intended to present a preliminary design of the data governance framework and tools that are being developed in the INFINITECH project. Particularly, a first version of the description of the tools is provided including some schemas and specifications. In each subsection a detailed technical design of the tools is provided, including the internal modules, its basic requirements, and the high-level design features. These designs will be refined during the forthcoming period, and their final version will be gathered in the next deliverable of task T3.5 (D3.13), which will be delivered in M22 (July 2021).

### 3.1 Data pseudonymization tool

For certain analytical tasks, data needs to be pseudonymized, which enables analysis at the level of the individual, but does not include the possibility of personal identification. Since different datasets include various levels of personal information, the pseudonymization process must include not only IDs, but also other data fields which may allow personal identification.

Since there is no single easy solution to pseudonymization that works for all approaches in all possible scenarios, a high level of competence is required in order to apply a robust pseudonymisation process, possibly reducing the threat of discrimination or re-identification attacks, while maintaining the degree of utility necessary for the processing of the pseudonymised data.

When and which tool or approach to use for data pseudonymization is highly dependent on the data flow of the research. However, it is clear enough that the procedure should be implemented as early as possible in the data processing flow, and as close to the source as possible. The range of database fields that need to be pseudonymized is highly dependent on how wide is the range of personal information in the selected database. Users have to take into consideration not only IDs but also identify implicit information, which could lead to identifying a particular individual.

A general approach for pseudonymization consists of replacing one attribute from a record (typically unique) by another. The natural person is therefore still likely to be identified indirectly. However, the process reduces the linkability of a dataset with the original identity of a data subject; as such, it is a useful security measure but not a method of anonymization. Note that the result of the pseudonymization procedure can be independent of the initial value (as is the case of a random number generated by the controller or a surname chosen by the data subject) or it can be derived from the original values of an attribute or set of attributes (e.g., through a hash function or encryption scheme).

Focusing on the pseudonymization tool that is being developed within the INFINITECH project, it is important to highlight that it will be implemented in a form of service. It will allow inputs in a form of structured data, following a conventional format, and provide pseudonymized output in the same format. **It will support pseudonymization of unique identifiers and generalization of numeric and time-stamp fields** in batch mode using a REST API. Figure 2 shows a schema of the service. Users would first provide configuration including:

- input data attributes, including types of data (ID, time stamp, numeric types),
- the corresponding level of pseudonymization and
- the required type of generalization for numeric and time stamp data.

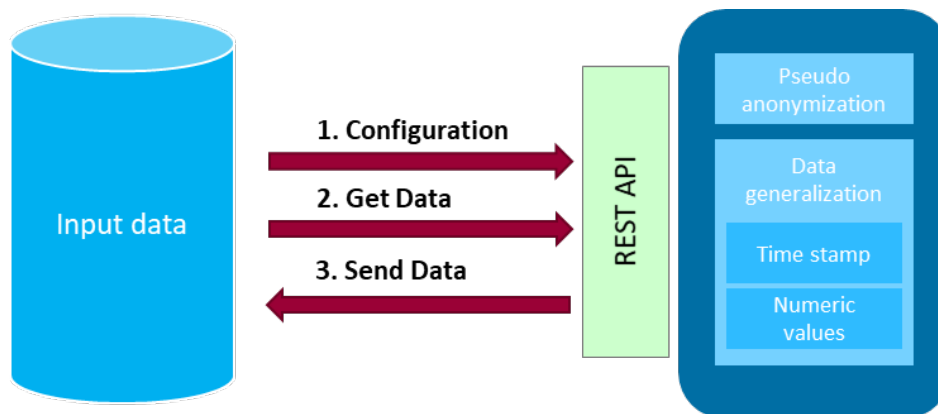


Figure 2 - Schema of the pseudonymization tool.

As it was mentioned above, the implemented tool will provide functionalities for **generalization** (the level needs to be specified in the configuration) of:

- numeric fields (rounding to different levels: integer number, tens, hundreds, thousands)
- time stamp data (rounding to the level of hour, day, week)

Regarding, the pseudonymization of **unique identifiers**, the following approaches are considered:

- Hash function (plus a nonce/salt) to generate a unique hexadecimal number of variable length, depending on the selected algorithm (e.g. MD5, SHA-2, SHA-3).
- Hashing with Key: A robust approach to generate pseudonyms is based on the use of keyed hash functions. The main difference from the conventional hash functions is that, for the same input (a data subject's identifier), several different pseudonyms can be produced, according to the choice of the specific key. Therefore, if the data controller needs to assign the same pseudonym to the same individual, then the same secret key should be used. Thus a secure keyed-hash function, with properly chosen parameters, is needed (such as HMAC with a secure hashing function). The secret key needs to be unpredictable and of sufficient length, e.g. 256 bits.

Hence, recalling the definition of pseudonymization in the GDPR, the data controller should keep the secret key securely stored separately from other data, as it constitutes the additional information, i.e. it provides the means for associating the individuals.

## 3.2 Data anonymization

The aim of the anonymization is to process personal data in order to irreversibly prevent identification. Through this mechanism, the data can be modified in many different ways and degrees and these modifications will change the privacy and utility of the dataset. In general, as the anonymization of the data increases so does their privacy but at the expense of their utility which decreases. Thus, there exists a trade-off between these two levels which must be defined by the final user when an anonymization procedure is required. The anonymization tool that is being developed within the INFINITECH project will allow the user to select the anonymization level that best fits the required privacy and utility.

The anonymization tool will be developed as a service that can be instantiated through a REST API (see Figure 3). This service will always require a **configuration** that indicates how the anonymization should be performed. In order to obtain this configuration, the client would need to send a request to the service REST API (1), indicating the parameters of the database that contains the non-anonymized data (such as the type of database, location, name of the table that contains the data, name and type of the data columns, and credentials). The anonymization service would then retrieve the data from the database (2, 3) and apply all the possible anonymization operations (generalization, randomization, deletion) to each of the data columns, calculating the corresponding values of the resultant privacy and utility metrics for each of the possible operations (4). Finally, all this information would be wrapped-up as a file and returned to the client (5) for its



use during the actual anonymization process that would take place later. Even though the act of obtaining this configuration for the anonymization service takes a relatively long time, it is only executed once.

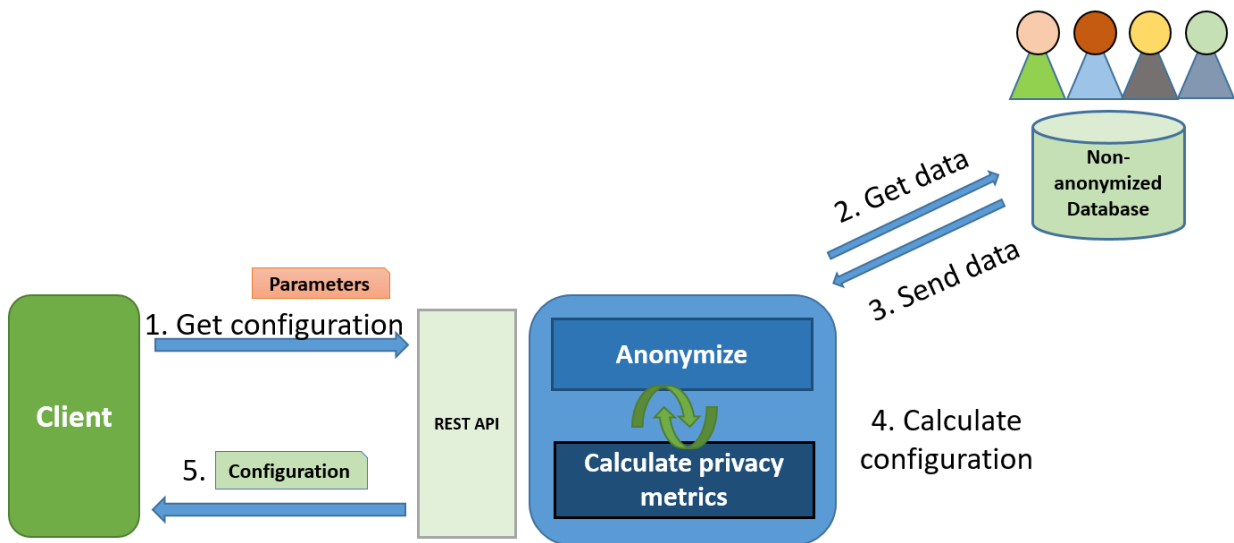


Figure 3 - Obtaining of the configuration for anonymizing a dataset.

The anonymization tool is intended to be used in two modes, batch or streaming. Batch processing is the running of a job that can be scheduled to run as resources permit or can run without end user interaction, i.e., the processing happens on blocks of data that have already been stored over a period of time. Thus, some type of storage is required to load the data (file, records, etc.). It is often used when you do not need real-time analytics results. Unlike batch processing, stream processing allows us to process data in real time as they arrive within a small time period. From the point of view of the performance, the latency of batch processing will be from minutes to hours while the latency of stream processing will be in seconds or milliseconds. It is important to note that although the batch option can take a long time in comparison with the streaming mode, it allows anonymization of all the data at once. These two modes of data processing are taken into account in order to provide a complete and flexible tool which is able to deal with different types of use cases, such as those considered in pilot 11 and 12. Both pilots try to improve the analysis, definition and assignment of risk profiles in car and health insurance, respectively, applying Artificial Intelligence technologies. Since the information collected for training the models is sensitive in both cases (e.g., GPS location of the connected car in pilot 12 and physical activity of the user in pilot 12) the anonymization tool will be used.

In the *batch mode* (see Figure 4), the service receives an asynchronous request from a client in order to anonymize a particular dataset (1,2). Then, the service retrieves the non-anonymized dataset from the selected database (3,4) and, once the data is anonymized using the configuration previously obtained, it calculates a set of privacy and utility metrics in order to measure the efficiency of the anonymization process (5). If the values of the privacy and utility metrics are above a certain threshold (defined by the client), then anonymized data is stored in the destination database (6), and the values of the anonymization privacy and utility metrics are sent back to the client, indicating that the anonymization process is completed (7). On the other hand, if the values of the privacy and utility metrics are under the expected threshold, the service selects automatically the following anonymization configuration and repeats the process (anonymizing the data and measuring the metrics) (5) until the threshold is finally overcome.

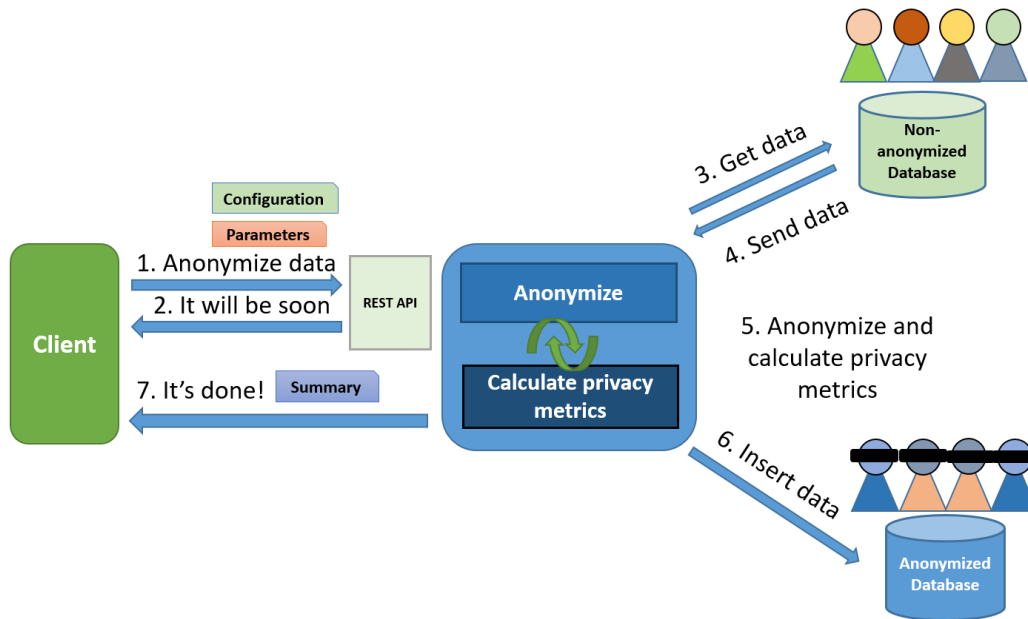


Figure 4 - Operation of the anonymization service in *batch mode*.

In order to perform an adequate anonymization, the initial request that is sent to the API REST in *batch mode* would require including the following information:

- **Parameters for the connection to the source database:** type of the database, location (URL), name of the source table, name of the columns, data type of the columns and credentials (with read-only permissions).
- **Parameters for the connection to the destination database:** type of the database, location (URL), name of the table, and credentials (in this case with write permissions).
- **Callback URL and authentication token:** since the anonymization procedure can take a long time, the operation is designed to work asynchronously. This implies that once the process has ended, the service notifies it to the client by sending a request (with the authentication token) to the specified callback URL.
- **Configuration file:** it includes details on the different anonymization operations that can be applied to each of the data columns.

Furthermore, the anonymization service can also perform operations under *streaming mode* (Figure 5). In order to do this, it is necessary to configure the input (for the non-anonymized data) and output (for the anonymized data) queues of the service, and assign them a configuration for the anonymization procedure of a particular type of data. If two clients want to anonymize the same data types, for example, GPS location and age, they will have to use different queues, because each client needs his/her particular configuration.

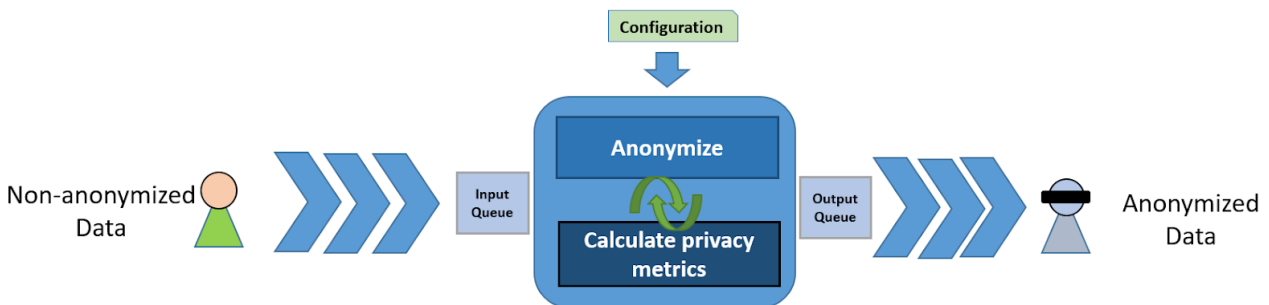


Figure 5 - Operation of the anonymization service in *streaming mode*.

The procedure of configuring the anonymization service in the *streaming mode* must be performed before performing any actual anonymization, by sending a request to the REST API of the service (1,2). This request should include the connection data for the input and output queues (for retrieving the non-anonymized data and sending its anonymized counterpart, respectively), as well as the configuration that will indicate how the data must be transformed (Figure 6).

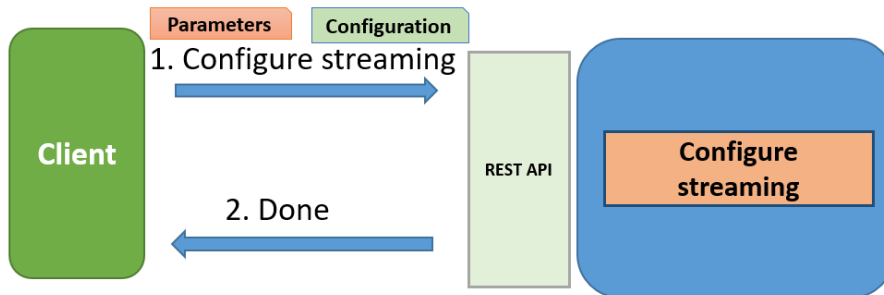


Figure 6 - Configuration of the *streaming mode* of the anonymization service.

### 3.3 Digital user onboarding services tool

Digital user onboarding systems will be used in the financial sector, including pilot 4, to let customers of the bank or Fintechs to create their own user identities that will be used later to access the services related with the corresponding application or portfolio. In the case of pilot 4, it will be the identity associated with the portfolio, which will be personalized for that identity.

INFINITECH onboarding system will be an adaptation of the DUOS (Digital User Onboarding System) developed in [46] for the ARIES project. **DUOS allows for remote user registration using eID or electronic password and provides multi-factor authentication combining images of the face of the user with the certificates stored in the eID or passport.** As these eID and passports are valid national identifiers, they will support and complement the SPeIDI mechanisms explained in section 3.4.

Figure 7 shows the overall work procedure of DUOS. During the **enrolment phase**, the user asks to create a new identity in the system (a virtual entity) and provides all needed authentication media for that (in practice a valid eID or passport contained in a chip), plus all required data about them that will be needed later to confirm their data. Then, the DUOS application returns the virtual identity, storing it in the device.

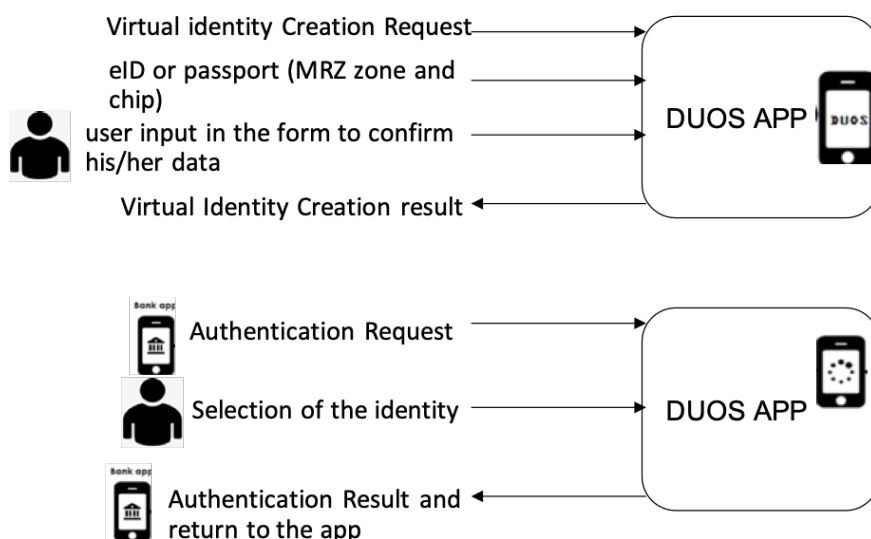


Figure 7 - Overall work procedure of DUOS

Once the user has successfully created their virtual identity with the DUOS application, this can be invoked by any application (which in our case will be the INFINITECH system). It will request the authentication to the DUOS app, which will allow the user to choose the proper virtual identity and return it to the requesting application.

Figure 8 shows the process of the **onboarding operation** for INFINITECH. When the user wants to register on the platform, they need to provide either their eDNI (national electronic identifier) or electronic passport, which has a chip that contains the credentials of the user. The onboarding tool of INFINITECH will read the chip and get a capture of the face of the person to compare them and verify them against eIDAS through SPeIDI. If the verification succeeds, the system will issue a new vID that will be able to be used later to access the system.

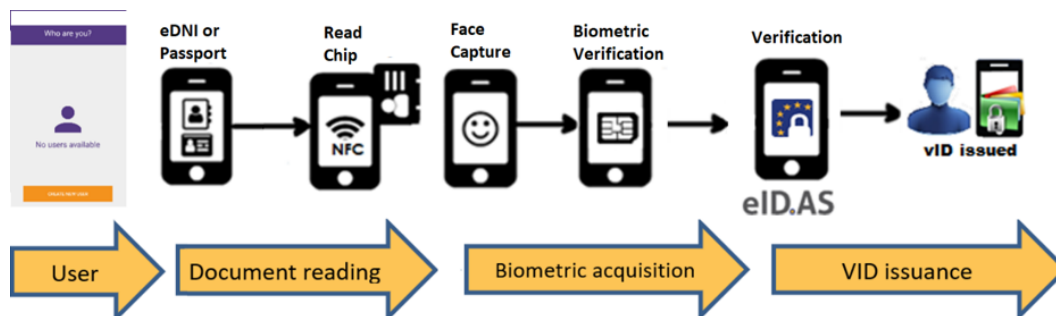


Figure 8 - INFINITECH Onboarding operation.

Figure 9 shows how the combined **authentication** of DUOS-SPeIDI works. The user accesses the bank application which asks for authentication, redirecting to the DUOS app. The app will select the vID stored in the device and take a video capture of the user for comparing the characteristic features of the captured face with the one included in the vID. Finally, if the face coincides, it redirects to SPeIDI to get the user additionally authenticated against eIDAS infrastructure.



Figure 9 - INFINITECH authentication against vID generated in the onboarding process.

### 3.4 Authentication tool for the eIDAS infrastructure

eIDAS, which stands for electronic IDentification, Authentication and trust Services, is a mandatory regulation from the European Union for **authentication and trust services in electronic transactions** in it. As such, it also covers authentication of both physical (citizens) and legal persons. As stated in [47], the Regulation provides the regulatory environment for the following main aspects related to electronic transactions:

- **Advanced electronic signing**, which (i) must be able to uniquely identify and link to the signatory, (ii) it must be created with electronic signature, (iii) it must be under sole control of the signatory and (iv) it must be linked to the data signed so any change in them is detectable.
- **Qualified electronic signatures and certificates**, which must ensure (i) the confidentiality of the electronic data used for signing, including protection against those data being used by others, (ii) that the signature can occur only once, (iii) that the data to be signed have not been altered, and that the data will not be prevented to be presented to the signator prior to signing.
- **Generation and managing of the creation of the signature** may only be done by a qualified trust service provider.
- **Duplicates of data signed** will be used only for backup purposes and kept to the minimum for the continuity of the service.

- **Trust services**, defined as electronic services that create, validate and verify electronic signatures, timestamps seals and certificates.

To comply with all these requirements, the project will provide a **solution for authentication compliant with eIDAS regulation** named SPeIDI (Service Provider for eIDAS Integration). The solution will support:

- Authentication for citizens, including those trying to access other country services different from their origin country.
- Compatibility with eIDAS Network for authentication purposes.
- Use of eID issued by European National authorities according to the EU eID schemas.
- Strong cross-border authentication using more secure credentials.
- UI easing usability and privacy, which includes: (i) user country selection which displays a list of mandatory and optional attributes requested by the SP and (ii) asks for user consent which provides privacy policy from the SP.

Figure 10 shows an example of how SPeIDI will work. The steps are described as follows:

1. User tries to access an INFINITECH system, which redirects him/her to the page of their national eIDAS service provider (in this case, this will be SPeIDI in JWT format).
2. SPeIDI will translate the request to the SAML2 language used by the eIDAS network and will forward it to the Spanish service provider through the eIDAS network.
3. the eIDAS network will provide a mapping service for SP attribute name/eIDAS attribute name and will issue a proper request to the Spanish eIDAS identity provider. In this case the provider will be the Spanish police, which, like any member state authorized issuer, had previously issued a national eID for the user.
4. The Spanish police will issue a page to the user so that he/she will be able to enter their national eID card and checks it.
5. Once the identity of the user is checked by the national identity provider, it replies the SAML2 request to the EIDAS network.
6. At this moment, the eIDAS network will reply to SPeIDI with the requested authentication token and attributes.
7. SPeIDI will translate the SAML request to JWT and will return the obtained identity and attributes to the INFINITECH application so they can be used in checking access to the different resources that it has.

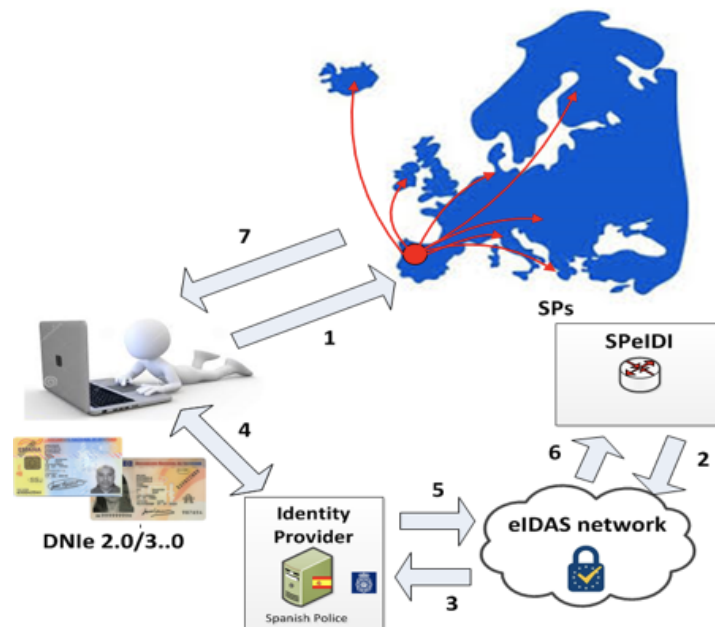


Figure 10 - Spanish citizen accessing online service provided by a company based in the EU and connected to eIDAS network.

## 4 Conclusions

This document reports the work that has been currently done in the scope of the task “T3.5 Data Governance Mechanisms” of the INFINITECH project whose goal is to implement and provide the following data governance building blocks: (i) a pseudonymization tool, (ii) a mechanism for anonymizing datasets, (iii) a mobile digital user onboarding services with virtual eID derived from government issued documents and (iv) a solution for authenticating citizens and/or businesses against the pan European eIDAS infrastructure.

Towards that goal, firstly a state of the art analysis has been made on the technologies included on each of the blocks. In order to ensure privacy, several pseudonymization methods can be applied to data to either pseudonymize direct identifiers or scramble the data. Counter and random number generator (RNG), cryptographic hash functions, symmetrically encrypted identifiers or data masking are examples of generating the pseudonyms out of raw identifiers. The pseudonymization block that is being developed within the project will support pseudonymization of unique identifiers and generalization of numeric and time-stamp fields. It will work in batch mode by using a REST API and a configuration file provided by the user including input data attributes, the corresponding level of pseudonymization and the required type of generalization for numeric and time stamp data will be necessary.

Data anonymization (or de-identification) tries to handle personal data in order to irreversibly prevent identification. Most of the anonymization techniques are based on two main methods: randomization and generalization methods. Noise addition, permutation techniques or differential privacy are examples of the first option, among others. Regarding the generalization of the information contained in the dataset, k-anonymity, l-diversity, t-closeness are other alternatives to anonymize data. One of the solutions that INFINITECH will provide is a tool for data anonymization that determines automatically the best anonymization configuration for each application. Different anonymization algorithms will be applied to avoid the appearances of data combinations that could lead to a possible re-identification of the data subjects. Additionally, we will calculate a set of privacy and utility metrics that allow measurement of the risk that remains after anonymizing the dataset and the impact of the anonymization process on the quality of the data.

Paying attention to the digital user onboarding, this is the process of enrolling new users ensuring that they can access all the services and products contracted in a remote and secure way. The procedure involves two technical challenges: authentication and authorization or access control. The INFINITECH onboarding system that is being developed will allow for remote user registration using eID or electronic password and provides multi-factor authentication combining images of the face of the user with the certificates stored in the eID or passport.

Finally, eIDAS (electronic Identity And trust Services) Regulation contributes to secure cross-border electronic transactions and central building blocks of the Digital Single Market. Within the project, a solution for authenticating citizens and/or businesses against the eIDAS infrastructure will be developed, named SPeIDI (Service Provider for eIDAs Integration). It will provide a cross-border strong authentication mechanism based on eIDs and will support authentication for citizens, compatibility with eIDAS Network, use of eID issued by European National authorities according to the EU eID schemas, strong cross-border authentication using more secure credentials and UI easing usability and privacy.

The result of the previous analysis has been used to carry out the preliminary design of each of the data governance tools that are being developed within the INFINITECH project. It is worth mentioning that this is the first report of the work to be done in the scope of T3.5, and there will be two more deliverables. Particularly, the next one, INFINITECH-D3.13, will include the final design of the tools and the advances related to the development of them and will be delivered in M22 (July 2021). Additionally, it is expected that these tools will start being validated through the specific pilots.

To conclude, the progress of the task T3.5 is in accordance with what has been planned.

## Appendix A: Literature

- [1] “Pseudonymization,” *Imperva*. <https://www.imperva.com/learn/data-security/pseudonymization/>.
- [2] European Union Agency for Cybersecurity, “Pseudonymisation Data protection GDPR Pseudonymisation techniques and best practices,” 2019, doi: 10.2824/247711.
- [3] J. Wang, Y. Luo, Y. Zhao, and J. Le, “A Survey on Privacy Preserving Data Mining,” in *2009 First International Workshop on Database Technology and Applications*, Apr. 2009, pp. 111–114.
- [4] Electronic Privacy Information Center, *Privacy and Human Rights 2005: An International Survey of Privacy Laws and Developments*. 2007.
- [5] “Article 29 Working Party. Opinion 05/2014 on Anonymisation Techniques (2014).” [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf) (accessed Sep. 14, 2020).
- [6] A. Narayanan and V. Shmatikov, “De-anonymizing Social Networks,” in *2009 30th IEEE Symposium on Security and Privacy*, May 2009, pp. 173–187.
- [7] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, “A Practical Attack to De-anonymize Social Network Users,” in *2010 IEEE Symposium on Security and Privacy*, May 2010, pp. 223–238.
- [8] A. Narayanan and V. Shmatikov, “Robust De-anonymization of Large Sparse Datasets,” in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, May 2008, pp. 111–125.
- [9] Y. Song, X. Lu, S. Nobari, S. Bressan, and P. Karras, “On the privacy and utility of anonymized social networks,” *International Journal of Adaptive, Resilient and Autonomic Systems (IJARAS)*, vol. 4, no. 2, pp. 1–34, 2013.
- [10] K. Mivule, “Utilizing Noise Addition for Data Privacy, an Overview,” *arXiv [cs.CR]*, Sep. 16, 2013.
- [11] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, “Aggregate Query Answering on Anonymized Tables,” in *2007 IEEE 23rd International Conference on Data Engineering*, Apr. 2007, pp. 116–125.
- [12] C. Dwork, A. Roth, and Others, “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [13] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” 1998, [Online]. Available: [http://epic.org/privacy/reidentification/Samarati\\_Sweeney\\_paper.pdf](http://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf).
- [14] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, “l-diversity: Privacy beyond k-anonymity,” *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 3–es, 2007.
- [15] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,” in *2007 IEEE 23rd International Conference on Data Engineering*, Apr. 2007, pp. 106–115.
- [16] A. R. Beresford and F. Stajano, “Location privacy in pervasive computing,” *IEEE Pervasive Comput.*, vol. 2, no. 1, pp. 46–55, Jan. 2003.
- [17] J. Krumm, “Inference Attacks on Location Tracks,” in *Pervasive Computing*, 2007, pp. 127–143.
- [18] M. Gruteser and B. Hoh, “On the Anonymity of Periodic Location Samples,” in *Security in Pervasive Computing*, 2005, pp. 179–192.
- [19] Baik Hoh and M. Gruteser, “Protecting Location Privacy Through Path Confusion,” in *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM’05)*, Sep. 2005, pp. 194–205.
- [20] Baik Hoh, M. Gruteser, Hui Xiong, and A. Alrabady, “Enhancing Security and Privacy in Traffic-Monitoring Systems,” *IEEE Pervasive Comput.*, vol. 5, no. 4, pp. 38–46, Oct. 2006.
- [21] Y.-A. de Montjoye, Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The privacy bounds of human mobility,” *Scientific Reports*, vol. 3, no. 1. 2013, doi: 10.1038/srep01376.
- [22] D. J. Patterson, L. Liao, D. Fox, and H. Kautz, “Inferring High-Level Behavior from Low-Level Sensors,” in *UbiComp 2003: Ubiquitous Computing*, 2003, pp. 73–89.
- [23] J. Froehlich and J. Krumm, “Route prediction from trip observations, Soc,” *Automot. Eng. Spec. Publ*, vol. 2193, p. 53, 2008.
- [24] J. Krumm and E. Horvitz, “Predestination: Inferring Destinations from Partial Trajectories,” in *UbiComp 2006: Ubiquitous Computing*, 2006, pp. 243–260.

- [25] C. Bettini, X. S. Wang, and S. Jajodia, "Protecting Privacy Against Location-Based Personal Identification," in *Secure Data Management*, 2005, pp. 185–199.
- [26] M. Duckham and L. Kulik, "Location privacy and location-aware computing," *Dynamic & mobile GIS: investigating change in*, 2006, [Online]. Available: [https://www.academia.edu/download/44974954/Location\\_privacy\\_and\\_location-aware\\_comp20160421-7519-143glj.pdf](https://www.academia.edu/download/44974954/Location_privacy_and_location-aware_comp20160421-7519-143glj.pdf).
- [27] M. Gruteser and D. Grunwald, "Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking," *Proceedings of the 1st international conference on Mobile systems, applications and services - MobiSys '03*. 2003, doi: 10.1145/1066116.1189037.
- [28] B. Gedik and L. Liu, "Location Privacy in Mobile Systems: A Personalized Anonymization Model," *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*. doi: 10.1109/icdcs.2005.48.
- [29] G. Zhong and U. Hengartner, "A distributed k-anonymity protocol for location privacy," *2009 IEEE International Conference on Pervasive Computing and Communications*. 2009, doi: 10.1109/percom.2009.4912774.
- [30] M. Duckham and L. Kulik, "A Formal Model of Obfuscation and Negotiation for Location Privacy," *Lecture Notes in Computer Science*. pp. 152–170, 2005, doi: 10.1007/11428572\_10.
- [31] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar, "Preserving User Location Privacy in Mobile Data Management Infrastructures," *Privacy Enhancing Technologies*. pp. 393–412, 2006, doi: 10.1007/11957454\_23.
- [32] C. A. Ardagna, M. Cremonini, E. Damiani, S. D. C. di Vimercati, and P. Samarati, "Location Privacy Protection Through Obfuscation-Based Techniques," *Data and Applications Security XXI*. pp. 47–60, 2007, doi: 10.1007/978-3-540-73538-0\_4.
- [33] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets Practice on the Map," *2008 IEEE 24th International Conference on Data Engineering*. 2008, doi: 10.1109/icde.2008.4497436.
- [34] S.-S. Ho and S. Ruan, "Differential privacy for location pattern mining," *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS - SPRINGL '11*. 2011, doi: 10.1145/2071880.2071884.
- [35] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," *Proceedings of the 2012 ACM conference on Computer and communications security - CCS '12*. 2012, doi: 10.1145/2382196.2382263.
- [36] R. Dewri, "Local Differential Perturbations: Location Privacy under Approximate Knowledge Attackers," *IEEE Transactions on Mobile Computing*, vol. 12, no. 12. pp. 2360–2372, 2013, doi: 10.1109/tmc.2012.208.
- [37] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability," *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13*. 2013, doi: 10.1145/2508859.2516735.
- [38] Martin David, Jorge Bernal, Julien Bringer, Nicolas Notario, Eduardo Gonzales, "D3.1 – ARIES eID ecosystem technical design." 2017.
- [39] "EUR-Lex - 32014R0910 - EN - EUR-Lex." [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L\\_.2014.257.01.0073.01.ENG](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.257.01.0073.01.ENG) (accessed Sep. 14, 2020).
- [40] A. Crespo, "PowerPoint presentation: eIDAS-Compliant Cross-Border Authentication - Alberto Crespo." 2018.
- [41] B Campbell C Mortimore, "Security Assertion Markup Language (SAML) 2.0 Profile for OAuth 2.0 Client Authentication and Authorization Grants." 2015, [Online]. Available: <https://tools.ietf.org/html/rfc7522>.
- [42] International Organization for Standardization, "ISO/IEC 29115:2013 Information technology — Security techniques — Entity authentication assurance framework." 2013, [Online]. Available: <https://www.iso.org/standard/45138.html>.
- [43] C. Gómez, "eID under eIDAS Building trust in a digital society - DG CONNECT-European Commission." [Online]. Available: [http://st.fbk.eu/sites/st.fbk.eu/files/20180316\\_eidas\\_oauth\\_security\\_workshop.pdf](http://st.fbk.eu/sites/st.fbk.eu/files/20180316_eidas_oauth_security_workshop.pdf).



- [44] “CEF Digital Home.”  
<https://ec.europa.eu/cefdigital/wiki/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home> (accessed Sep. 14, 2020).
- [45] K. K. Petros Kavassalis, “D4.3 Operational and Technical Documentation of SP (ATHEX, Hellenic Post) integration (production).” 2018, [Online]. Available: <http://www.leps-project.eu/sites/default/files/leps/public/content-files/deliverables/LEPS%20D4.3%20Operational%20and%20Technical%20Documentation%20of%20SP%20integration.pdf>.
- [46] ARIES consortium, “D4.1 ARIES prototype instantiation.” 2018, [Online]. Available: <https://www.aries-project.eu/sites/default/files/aries/public/content-files/deliverables/D4.1%20ARIES%20prototype%20instantiation.pdf>.
- [47] J. Dumortier, “REGULATION (EU) NO 910/2014 ON ELECTRONIC IDENTIFICATION AND TRUST SERVICES FOR ELECTRONIC TRANSACTIONS IN THE INTERNAL MARKET (EIDAS REGULATION),” in *EU Regulation of E-Commerce*, Edward Elgar Publishing, 2017.